Was gehört in ein nationales Gesprächskorpus?

Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim)

Arnulf Deppermann und Martin Hartung

Inhalt

1. Warum ein nationales Gesprächskorpus?
2. Gegenstandsbestimmung
3. Theoretische Kriterien der Stratifikation
3.1 Merkmale des Sprechereignisses
3.2 Merkmale der Sprecher
3.3 Merkmale der Sprache der Interaktion
4. Stratifikationskonzepte bestehender Korpora
4.1 British National Corpus (BNC)
4.2 Corpus Gesproken Nederlands (CGN)
4.3 Göteborg Spoken Language Corpus (GSLC)
4.4 Evaluation der Stratifikationskonzepte
5. Theoretische Probleme des Korpusdesigns
5.1 Repräsentativitätsproblem
5.2 Probleme des konstitutionstheoretischen Zugriffs
5.3 Probleme der Parametrisierung
6. Praktische und forschungsethische Probleme der Korpusstratifikation
6.1 Rechtliche und ethische Anforderungen
6.2 Probleme der Erhebbarkeit von Daten
6.3 Probleme verfügbarer und repräsentierbarer Datenqualität
7. Praktische Prioritäten des Korpusaufbaus
8. Ausblick
9. Literaturverzeichnis
10. Internetquellen der im Text aufgeführten Korpora

1. Warum ein nationales Gesprächskorpus?

In den vergangenen vierzig Jahren wurden von GesprächsforscherInnen im deutschen Sprachraum eine große Menge von Gesprächsdaten in unterschiedlichsten gesellschaftlichen Interaktionssituationen erhoben (s. GLAS/EHLICH 2000; WAGENER/BAUSCH 1997). Diese Daten wurden in der Regel mit öffentlichen Mitteln (der Universitäten und der Forschungsförderinstitutionen) im Rahmen von empirischen Forschungsprojekten gesammelt, (zum Teil) aufbereitet (dokumentiert, transkribiert, digitalisiert) und (zum noch geringeren Teil) ausgewertet. Doch „bleibt (...) der Normalfall der, dass gesprächsanalytische Daten, nachdem sie erstellt und in einem relativ begrenzten Zusammenhang analysiert wurden, der wissenschaftlichen Öffentlichkeit nicht, oder zumindest nicht in einer brauchbaren Form, zur Verfügung stehen, ja dass größtenteils sogar nicht einmal verlässliche Informationen über ihre Existenz und Zusammensetzung erhältlich sind" (SCHMIDT 2005: 106). Diese Situation ist in mehreren Hinsichten unbefriedigend (vgl. SCHMIDT 2005: 107):

- Unter *ökonomischen* Gesichtspunkten werden durch die nicht gegebene Möglichkeit zur Nachnutzung von Daten aus Forschungsprojekten in jedem neuen Projekt wieder die gleichen, überaus zeitintensiven Arbeitsschritte des Gewinns des Feldzugangs, der Datenerhebung und der Datenaufbereitung notwendig. Diese Arbeitsschritte machen in der Gesprächsforschung etwa im Vergleich zur Text- oder Medienforschung einen ungleich höheren Aufwand aus. Je nach Schwierigkeit des Feldzugangs, Umfang der Dokumentation und Zusatzmaterialien, Komplexität des Datenmaterials (Audio vs. Video, dyadische vs. Mehrparteiengespräche) und Transkriptionsgenauigkeit kann man von einem Quotienten von ca. 1:50 bis 1:150 Zeitaufwand für die Korpuserstellung pro Aufnahmeminute ausgehen, bevor überhaupt mit der Analyse begonnen werden kann.
- Unter *quantitativen* Gesichtspunkten führt die fehlende Möglichkeit der Nachnutzung vorhandener Korpora dazu, dass der wissenschaftlichen Gemeinschaft keine sich kumulativ entwickelnde Datenbasis zur Verfügung steht. Es ist somit weder möglich, bereits vorliegende Korpora für Sekundäranalysen (andere Auswertungen als im ursprünglichen Forschungsvorhaben) zu nutzen, noch mehrere Korpora in Bezug auf eine Fragestellung vergleichend zu untersuchen (Metaanalysen) noch die Befunde in Bezug auf ein eigenes, neu erhobenes Korpus kontrastiv oder ergänzend durch den Vergleich mit den Gegebenheiten in bestehenden Korpora zu profilieren.

- Unter *didaktischen* Gesichtspunkten führt die Unzugänglichkeit von Aufnahmen authentischer Kommunikationspraxis dazu, dass die universitäre Lehre sowohl im Bereich der Gesprächsforschung selbst bzw. aller wissenschaftlicher Fragestellungen, die sich für authentische Gespräche interessieren (z.B. Gesprochene-Sprache-Forschung, Phonetik, Sprachsoziologie, Psycholinguistik, Unterrichtsforschung), sowie im Bereich der Vermittlung von Deutsch als Fremd- bzw. Zweitsprache nur in sehr beschränktem Maße auf empirische Daten zurückgreifen kann.

Zusammengenommen führen diese Probleme dazu, dass sich die Gesprächsforschung als Richtung innerhalb der Linguistik zeitlich wie qualitativ, d. h. hinsichtlich der Reichweite und Differenziertheit der Ergebnisse, erheblich schleppender weiterentwickelt als dies sein könnte. Dies ist gerade im Vergleich zum Boom der textgestützten Korpuslinguistik evident. Ihr Erfolg beruht essentiell darauf, dass mittlerweile öffentlich verfügbare, große und ständig wachsende Textkorpora zur Verfügung stehen. Im deutschen Sprachraum sind dies v.a. das Deutsche Referenzkorpus am Institut für Deutsche Sprache (http://www.ids-mannheim.de/kl/projekte/korpora/), das über COSMAS II zugänglich ist (http://www.ids-mannheim.de/cosmas2/), die Korpora des Deutschen Wörterbuch der Gegenwartssprache (DWDS) der Berlin Brandenburgischen Akademie der Wissenschaften http://www.dwds.de/textbasis), sowie die Korpussammlung der Universität Leipzig (http://corpora.informatik.uni-leipzig.de/download.html). In der Gesprächsforschung müssen sich neue Forschungsvorhaben dagegen in der Regel zunächst einmal dem Korpusaufbau widmen. Nur in den seltensten Fällen werden korpusübergreifende Untersuchungen vorgenommen, die weitergehende Aussagen über die gesellschaftliche Verbreitung, die kontextgebundene Spezifität der Ausprägungen und die situative Adaptation von sprachlich-kommunikativen Phänomenen erlauben.

Nun gibt es für diesen Zustand viele gute und auch einige schlechte Gründe. Sie liegen nicht nur im Unwillen der ForscherInnen, potenziellen KonkurrentInnen die Früchte ihrer Arbeit ohne Gegenleistung verfügbar zu machen und damit ihnen selbst gegenüber einen Wettbewerbsvorteil zu verschaffen. Weitere Ursache sind in den besonderen methodologischen, technischen und rechtlich-ethischen Gegebenheiten der Gesprächsforschung zu suchen, welche die Weitergabe und Nachnutzung von Daten in einer Weise erschweren wie dies bei Schriftkorpora nicht der Fall ist (s. dazu SCHMIDT 2005). Diese Gründe sorgen dafür, dass der Versuch, bestehende Korpora zu sammeln und öffentlich zugänglich zu machen, vor sehr vielen Hürden steht, die teilweise gar nicht (rechtliche Restriktionen der Weitergabe und Nutzung von

Korpora, fehlende Dokumentationen) bzw. nur mit unverhältnismäßigem Aufwand (Digitalisierung, technische Nachbereitung, Vereinheitlichung von Datenformaten, Retranskription) zu überwinden sind. Entsprechend rar sind öffentlich verfügbare Korpussammlungen und erst recht übers Internet zugängliche Archive (vgl. MERKEL/SCHMIDT 2009; MCCARTHY/O'KEEFE 2008; WICHMAN 2008). In Deutschland sind hier lediglich die Datenbank Gesprochenes Deutsch (DGD: http://agd.ids-mannheim.de/html/dgd.shtml) des Instituts für Deutsche Sprache (s. FIEHLER/WAGENER 2005) zu nennen sowie für den Bereich Mehrsprachigkeit einige Bestände des Hamburger SFB 538 (http://www.exmaralda.org/corpora/sfbkorpora.html). Doch auch diese Sammlungen bieten aufgrund der Heterogenität der Art der Datenaufbereitung in den jeweiligen Ursprungsprojekten und ihrer thematischen Begrenzungen bzw. teils auch wegen des Alters der verfügbaren Daten nur sehr eingeschränkte Forschungsmöglichkeiten.

Aufgrund dieser unbefriedigenden Situation haben wir im Jahre 2008 am IDS damit begonnen, ein nationales Gesprächskorpus aufzubauen, das den „kommunikativen Haushalt" (LUCKMANN 1986) der deutschsprachigen mündlichen Kommunikationspraxis in seinen wesentlichen Ausprägungen repräsentieren soll (s. Abschn.2). Das regulative Ziel ist es, das volle Spektrum der privaten, institutionellen, öffentlichen und massenmedialen Anlässe und Typen mündlicher Kommunikation im Deutschen nach und nach durch Audio- und Videodatenaufnahmen zu dokumentieren, zu transkribieren und soweit als möglich der wissenschaftlichen Gemeinschaft zur Verwendung für Forschungs- und Lehrzwecke zur Verfügung zu stellen. Dementsprechend nennen wir dieses Korpus „Forschungs- und Lehrkorpus gesprochenes Deutsch" (FOLK: http://agd.ids-mannheim.de/html/folk.shtml). Beim Aufbau eines solchen Korpus kann man natürlich auf Kenntnisse und Erfahrungen aus der gesprächsanalytischen, korpuslinguistischen und texttechnologischen Arbeit zurückgreifen. An vielen Stellen betritt man aber auch Neuland. So sind z.B. Fragen zu klären und Probleme zu lösen, die die Wahl eines geeigneten Transkriptionssystems, die Konzeptualisierung von Metadatendeskriptoren, die Standardisierung von Datenformaten, die Implementierung geeigneter Such-, Anzeige- und Auswertungsfunktionen, die datenschutzrechtliche Behandlung von Korpusbeständen oder den gestuften Zugang unterschiedlicher Nutzergruppen zu unterschiedlichen Datenbeständen betreffen. Diese überaus wichtigen und für den Korpusaufbau zentralen Fragen, die Gegenstände unserer gegenwärtigen Entwicklungsarbeit sind, werden wir in den kommenden Jahren an anderer Stelle diskutieren. In diesem Text behandeln wir die mindestens ebenso zentrale Frage, wie eigentlich ein

solches nationales Korpus zusammengesetzt sein soll. Es geht also darum, nach welchen Kriterien entschieden werden soll, welche Ausschnitte der gesellschaftlichen Gesprächspraxis im Korpus enthalten sein sollen und wie das Korpus sukzessive zu erweitern ist. Mit anderen Worten muss die Datenstratifikation eines solchen Korpus theoretisch fundiert sein, will es denn den Anspruch erfüllen, ein „ausgewogenes Korpus" zu sein (LEMNITZER/ZINSMEISTER 2006). Dazu müssen wir zu Kriterien für Ausgewogenheit kommen, die auf den Erkenntnissen der Wissenschaften fußen, welche sich mit gesprochener Sprache, ihrer kommunikativen Praxis in sozialen Interaktionen und ihrer sozialen Variation befassen (s. Abschn. 3). Eine entsprechende Systematik der Korpusstratifikation ist nötig, um sowohl die angestrebte qualitative Zusammensetzung der „Vollversion" des Korpus, als auch seine quantitative Zusammensetzung zu bestimmen. Sie ist aber auch grundlegend für die Frage, wie die Metadatendokumentation und die Konzeptualisierung der Such- und Auswertungsfunktionalitäten des Korpus aussehen soll, da die Stratifikationskriterien zumindest potenziell auch dokumentations- und auswertungsrelevant sind. Hier ist der Vergleich mit ähnlichen Vorhaben hilfreich (Abschn. 4). Dabei werden einige grundlegende theoretische Probleme sichtbar, deren Lösung bisher nicht evident ist (Abschn. 5). So unabdingbar die theoretische Fundierung der Korpusstratifikation ist, so sehr kann sie gleichzeitig nur eine langfristig anzustrebende Zielperspektive sein, die in vielen Punkten leider aus prinzipiellen wie pragmatischen Gründen nicht vollständig umzusetzen sein wird. Viele praktische Hindernisse engen den Ausschnitt der gesellschaftlichen Kommunikationspraxis, die faktisch in einem Korpus repräsentiert werden kann, massiv ein (Abschn. 6). Insofern ist es erforderlich, Entwicklungsprioritäten zu definieren, die darauf abzielen, in möglichst kurzer Zeit, mit begrenzten, möglichst effizient einzusetzenden Mitteln ein möglichst großes, eine breite Nutzergemeinde interessierendes Korpus zu erstellen (Abschn. 7).

2. Gegenstandsbestimmung

Die Zielsetzung eines nationalen Gesprächskorpus scheint auf den ersten Blick klar vorzugeben, welche Arten sprachlicher Kommunikation das Korpus enthalten muss, nämlich fokussierte Interaktionen (i.S. von GOFFMAN 1963) im deutschen Sprachraum, in denen Sprache das Leitmedium der Kommunikation ist. Diese Definition muss aber in vielerlei Hinsicht präzisiert werden: in Bezug auf den zugrunde gelegten Gesprächsbegriff, das relevante Verständnis von „deutschem Sprachraum" und hinsichtlich der Frage, welche Erwartungen an eine

systematische Abdeckung sprachlicher und kommunikativer Variation zu stellen sind, wenn man die ins Auge gefassten Nutzergruppen des Korpus und ihre Nutzungsinteressen in Rechnung stellt. Das zu erstellende Korpus richtet sich in erster Linie an die Gesprächsforschung und die Gesprochene-Sprache-Forschung sowie an ForscherInnen mit einem Interesse an der Untersuchung authentischer Gespräche in Sozio-, Variations- und Medienlinguistik, Sprachsoziologie, Pädagogik und Kulturwissenschaft. Auch für ForscherInnen aus Psycholinguistik, Phonetik und Sprachtechnologie kann dieses Korpus hilfreich sein. Allerdings soll nicht angestrebt werden, sich an den in diesen Disziplinen vorherrschenden methodologischen Standards zu orientieren, die teilweise sehr anders als im Bereich der Gesprächsforschung und Gesprochene-Sprache-Forschung sind (z.B. phonetische Transkription, Erhebung unter standardisierten Bedingungen, Studio-Tonqualität etc.). Das Korpus soll außerdem auch Anschauungs- und Lehrmaterial zum heutigen gesprochenen Deutsch für Unterrichtszwecke bieten, insbesondere für die germanistische Hochschullehre und den DaF-Bereich.
Korpusdaten werden für Nutzer in grundsätzlich zwei Formen zugänglich sein:

- als Transkripte und – soweit dies datenschutzrechtlich möglich ist – Ton- bzw. Videoaufnahmen von Gesamtgesprächen, wobei die Tonaufnahmen mit dem Transkript text-ton-aligniert sind,
- als KWIC (*keyword in context*)-Ansichten von Treffern, die in Bezug auf Suchanfragen gefunden wurden.

Suchen können grundsätzlich an Transkriptmerkmalen (sprachliche Formen (Zeichenkette, mit regulären Ausdrücken), paraverbale Phänomene wie Pausen, gesprächsstrukturellen Merkmalen wie Sprecherwechsel, Überlappung, Turnbeginn/-ende) und an Metadaten zu Interaktionsbeteiligten (z.B. soziodemographische Merkmale, Sprachkenntnisse etc.) und Interaktionsereignissen (z.B. Interaktionstyp/Gattung, Anzahl der Beteiligten) ansetzen. FOLK bietet damit sowohl Materialien, die für eher einzelfallanalytisch orientierte Gesprächsanalysen von komplexen Interaktionstypen geeignet sind, als auch eine Grundlage für die Bildung von Kollektionen einer konversationellen Praktik. Durch den Bezug auf Metadaten können auch komparative Analysen durchgeführt werden, die den Zusammenhang zwischen Gesprächsphänomenen und gesprächsexternen Parametern bzw. interaktionstypologischen Merkmalen erkunden.

In Bezug auf seine Zusammensetzung bestimmen wir FOLK durch folgende Abgrenzungen näher:

- FOLK geht von einem *medial mündlichen Gesprächsbegriff* aus (vgl. KOCH/OESTERREICHER 1985) und beschränkt sich entsprechend auf mündliche Kommunikationsereignisse. Konzeptionell mündliche, aber medial schriftliche Daten wie Chat, SMS, Email werden nicht aufgenommen (vgl. THALER 2007).
- Es werden nur *gesprochene Daten* aufgenommen, keine Daten, die maschinell, per Sprachsynthese, generiert wurden.
- FOLK umfasst *authentische Interaktionen* im Sinne der Konversationsanalyse (vgl. DEPPERMANN 2008; SACKS 1984). Es werden keine nur für Forschungszwecke inszenierte Interaktionen („Rollenspiele") aufgenommen. Inszenierte Interaktionen werden dann aufgenommen, wenn die Inszenierung Teil der nicht-wissenschaftlichen Handlungspraxis ist (z.B. Theateraufführung, medienpädagogische Projektarbeit, Rollenspiele zu Fort- und Weiterbildungszwecken).
- FOLK nimmt *Interaktionstypen* auf. Ausschließlich monologische Kommunikationstypen (wie TV-Nachrichten, Predigten, Vorträge ohne Publikumsreaktion, Vorlesesprache) sind kein Bestandteil. Allerdings können auch Kommunikationsereignisse mit recht geringem Anteil an Interaktivität aufgenommen werden (wie z.B. Stadtführungen oder Seminarreferate) bzw. solche, die aus langen, aber aufeinander bezogenen Einzelbeiträgen bestehen (z.B. Parlamentsdebatten).
- Berücksichtigt werden *standardnahe bis gemäßigt dialektale Varianten* des gesprochenen Deutsch. Dialektale Daten werden nicht systematisch einbezogen. Die regionale Variation wird nicht systematisch erfasst, da dies allein schon aus Umfangsgründen utopisch wäre: Es müsste die areale Variationsbreite jedes erfassten Interaktionstyp dokumentiert werden! Insofern ist das Korpus nicht für die systematische kontrastive Untersuchung arealer Varianten geeignet. Wohl aber wird areale Variation in den Daten unsystematisch enthalten und (mithilfe entsprechender variationslinguistischer Dokumentation) auch punktuell untersuchbar sein, wie sich dies aus den nach anderen Stratifikationskriterien (Interaktionstyp, soziodemographischen Merkmalen etc.) im Korpus enthaltenen Daten ergibt.
- FOLK erfasst gesprochenes Deutsch. *Nicht-deutsche Sprachen* werden miterfasst, soweit sie in Situationen des Sprachkontakts bzw. der Sprachalternation mit dem Deutschen vorkommen. Ebenso werden *ethnolektale Varietäten* des Deutschen miterfasst. Auch hier wird aber keine systematische Erfassung in Bezug auf unterschiedliche Interaktionstypen angestrebt.

- FOLK beschränkt sich auf Aufnahmen aus dem Sprachraum, in dem *Deutsch Landessprache* ist. Auslandsdeutsche Varietäten werden nicht mit aufgenommen.
- FOLK umfasst Aufnahmen von Sprechern, die Deutsch als Erstsprache oder als Zweitsprache benutzen. *Lernervarietäten* werden nicht systematisch erfasst. Daten aus früheren Phasen der *Sprachentwicklung* (vor der Erwachsenensprache) sollen dagegen mit erfasst werden, da eine Erfassung des Sprechens unterschiedlicher Altersgruppen eine wichtige Stratifikationsdimension darstellt und da Interaktionen mit Kindern einen wichtigen Ausschnitt der gesellschaftlichen Kommunikationspraxis bilden.
- FOLK erfasst Daten von *nicht-sprachgestörten Sprechergruppen.* Sprachpathologien (wie Aphasie, Dysarthrie, Stottern) werden nicht systematisch erhoben.
- FOLK ist ein *synchrones Korpus*. Frühere Sprachstufen werden nicht berücksichtigt. Im Laufe seines Ausbaus wird FOLK jedoch eine diachrone Dimension gewinnen. Es muss dann entschieden werden, ob und wann Daten nacherhoben werden, da die entsprechenden Korpusbestandteile nicht mehr als Repräsentation der aktuellen Sprachwirklichkeit in den jeweiligen Interaktionstypen gelten können. Das Korpus wird auf jeden Fall auf die Dauer auch zu einem alltagskulturellen Archiv werden, das mittlerweile vergangene mündliche Kommunikationsformen für die Nachwelt bewahrt.
- In einer ersten Ausbaustufe beschränkt sich das Korpus auf die Einstellung von Audiodaten. *Videodaten* sollen in einer späteren Stufe hinzukommen. Empraktische Kommunikation kann zwar im Korpus aufgenommen werden. Sie hat jedoch keine Priorität, da ihre Untersuchbarkeit ohne Videodaten sehr eingeschränkt ist. Im Vordergrund stehen daher *verbal zentrierte Interaktionen.*

3. Theoretische Kriterien der Stratifikation

Für die Stratifikation des Korpus muss hinsichtlich folgender Fragen eine theoretische Basis geschaffen werden:

- Wie sind Erkenntnisse über die Konstitution von verbaler Interaktion sowie über die Verteilung und den Gebrauch von Sprachvariation in ein *taxonomisches, stratifikationsleitendes Konzept* umzusetzen?

- Nach welchen *sprecher- und ereignisbezogenen Kriterien* ist das Korpus zusammenzustellen? Wie sind sie in der Datenbank zu repräsentieren (Metadaten, Suchmöglichkeiten)? Welche variierenden Aspekte der sprachlich-kommunikativen Praxis sind zu berücksichtigen?
- Welchen *methodologischen Kriterien* muss die Repräsentation kommunikativer Wirklichkeit im Korpus genügen?
- Wie ist ein *ausgewogenes, „repräsentatives" Korpus* sicher zu stellen? Welche Datentypen sollen welches Gewicht im Korpus einnehmen?

Das zu erarbeitende Modell der Datenstratifikation beruht auf drei Klassen von Parametern, die systematisch variiert werden, um ein ausgewogenes und damit qualitativ repräsentatives Korpus zusammenzustellen:

- Merkmale des *Sprechereignisses*
- Merkmale der *Sprecher*
- Merkmale der *Sprache*

3.1 Merkmale des Sprechereignisses

Die Stratifikation nach Merkmalen des Sprechereignisses strebt an, die Variation der kommunikativen Handlungspraxis nach kommunikations- bzw. interaktionstheoretischen Kriterien zu erfassen. Um kommunikative Variation zu erfassen, kann man zwei Ausgangspunkte wählen: eine parametrisierte und eine Gattungssystematik.

Eine *parametrisierte Systematik* geht von Parametern aus, deren Ausprägung für die Konstitution unterschiedlicher Formen von Kommunikationsereignissen grundlegend ist. Hier kann angeknüpft werden an unterschiedliche Vorstellungen zu Merkmalen der Gesprächssituation (vgl. DEPPERMANN/SPRANZ-FOGASY 2001), wie sie z.B. im Freiburger Redekonstellationsmodell (STEGER et al. 1974) und dann weiterentwickelt in HENNE/REHBOCK (1982) vorgelegt werden. Merkmale von Kommunikationsereignissen werden hier jeweils durch Kontinua bzw. Dichotomien repräsentiert. Die Parameter, nach denen Kommunikationsereignisse unterschieden werden, werden hier als „Merkmale der Gesprächssituation" bezeichnet und nicht wie bei HENNE/REHBOCK (1982: 32ff.) als Kategorien von „Gesprächstypen". Der Grund besteht darin, dass als Stratifikationsprinzipien nur diejenigen Merkmale von Gesprächen tauglich sind, die einigermaßen statisch bzw. apriorisch vorhersehbar sind (s. u.). Für die Systematisierung der Stratifikation von FOLK codieren wir Gattungen (s. u.) nach folgenden Parametern:

- *Gesellschaftlicher Sektor*: Gesellschaftlicher Handlungsbereich, dem die Gattung zugehört: Bildung, Wirtschaft, Verwaltung, Recht, Medizin, Freizeit, Religion, Politik, Kunst, Medien
- *Ort*: Bindung des Gesprächstyps an bestimmte Örtlichkeit; z.B. Predigt in der Kirche, Familientischgespräch in privater Wohnung, Gerichtsverhandlung im Gericht
- *Zeit*: Bindung des Gesprächstyps an Tageszeit, Wochentag, Jahreszeit; z.B.: Frühstückstischgespräch, Montagmorgenkreis in Grundschule, Neujahrsansprache des Bundespräsidenten
- *Zugänglichkeit für Teilnehmer*: geschlossen (lebensweltliche Intimität/Bekanntschaft/Verwandtschaft oder institutionelle Rollenspezifikationen bzw. organisationale Mitgliedschaft als Zugangsvoraussetzung) vs. (teil-)öffentlich (unbeschränkter oder nicht rollengebunden beschränkter Zugang), z.B.: Bettgeflüster, Beratungsgespräch (geschlossen); Fakultätsversammlung (institutionell teilöffentlich); Wahlkampfrede (öffentlich)
- *Institutionalität*: Bindung der Interaktion an eine Institution; damit einhergehend institutionenbezogene Rollen der Gesprächsbeteiligten, z.B.: Berater und Ratsuchende im Beratungsgespräch
- *Mediale Realisierung*: technische Übertragungsmedialität, z.B.: Face-to-face-Gespräch, Telefongespräch, Bildtelefongespräch, Talk-Show (innerer Kommunikationskreis: *face to face*; äußerer Kommunikationskreis: massenmedial), TV-Nachrichten (massenmedial)
- *Anzahl der Teilnehmer*: dyadisch vs. Mehrpersonengespräch, z.B.: Psychoanalyse (dyadisch), Schlichtungsverhandlung (triadisch), Gruppentherapie (Mehrpersonengespräch)
- *Publikum*: Gespräche mit nicht bzw. nicht primär verbal aktiven Teilnehmern, z.B.: Talk-Show: Mehrpersonengespräch mit gestuftem Publikum: Talkgäste vs. Studiogäste vs. TV-Zuschauer
- *Sprecherwechsel*: dialogische vs. monologische Kommunikation
- *Vertrautheit der Teilnehmer*: unbekannt (Erstkontakt), bekannt, vertraut (Freundschaft, Familienmitgliedschaft)

- *Gesprächszweck*: Aufgabenbestimmung des Gesprächs in institutioneller Kommunikation; Handlungstypik in privater und Freizeitkommunikation, z.B.: Anamnesegespräch (Medizin); Beratungsgespräch (Psychologie); Dissen (Jugendkommunikation); Klagen (Familientischgespräch)
- *Soziale Rollen*: Beteiligungsrechte und -pflichten der Teilnehmer gemäß ihrer offiziellen Identitäten, die konstitutiv für ihre Zulassung zu einem privaten Kontext sind bzw. aufgrund derer sie an einer institutionellen Interaktion teilnehmen; nicht gemeint sind Beteiligungsrollen, die erst durch Gesprächsaktivitäten hergestellt werden wie z.B. Klagender-Tröster, Erzähler-Zuhörer, Freund-Freund in einem Arzt-Patient-Gespräch z.B. Mutter, Kind in Familientischgesprächen; Cliquenmitglied in Jugendkommunikation; Richter-Angeklagter-Zeuge-Protokollant-Rechtsanwalt in Gerichtsverhandlung
- *Themenvorgabe*: themen- vs. themenbereichsfixierte Gespräche vs. nicht-themenfixierte Gespräche. Der Grad der Fixierung bezieht sich dabei auf den Gesprächstyp, nicht auf das individuelle Gespräch. z.B.: themenfixiert: politische TV-Diskussion, Gerichtsverhandlung; themenbereichsfixiert: Beichte, Arzt-Patient-Gespräch, Talk-Show; nicht fixiert: Familientischgespräch, Kneipengespräch
- *Zeitliche Vorgabe*: begrenzte Dauer vs. unbegrenzte bzw. nicht vorgegebene Dauer; institutionelle und öffentliche Gespräche haben fast immer vorgegebene Zeitbegrenzungen (Sendezeiten, Sprechstunden, Sitzungszeiten etc.)
- *Vorbereitung*: spontan ohne Vorbereitung; vorbereitet; vorformuliert; vorgelesen. Der Vorbereitungsgrad kann für einzelne Teilnehmer unterschiedlich sein, z.B. beim Interview. Z.B.: spontan: Partygespräch; vorbereitet: Bewerbungsgespräch, freie Rede; vorformuliert: Gebet; vorgelesen: Nachrichten
- *Empraktischer Bezug*: Empraktisch sind solche Gespräche, in denen das Sprechen entweder nicht im Fokus der Aktivität steht, sondern nur eine ergänzende, organisierende oder komplementäre Rolle spielt, oder in denen verbale und nichtverbale, gegenständliche Handlungen eng miteinander verwoben sind.

Grafik 1 zeigt das von uns für die Planung von FOLK zugrunde gelegte Parameterset an einem Beispiel.

Gattung	Anmoderation
Lebensbereich	Fernsehen
Sektor	Medien
Ort	Studio
Zeit	unbestimmt
Zugang	● geschlossen ○ öffentlich ○ halböffentlich ○ divers
Institutionalität	● institutionell ○ nicht-institutionell
Medium	Fernsehen
Teilnehmerzahl	● eins ○ drei ○ >10 ○ zwei ○ 4-10
Publikum	● ja ○ nein
Sprecherwechsel	● monologisch ○ dialogisch
Vertrautheit	○ unbekannt ○ vertraut ● divers ○ bekannt ○ gemischt
Gesprächszweck	Anmoderation
Soziale Rollen	Moderator
Themenvorgabe	● themenfixiert ○ nicht-fixiert ○ bereichsfixiert
Zeitliche Dauer	● Vorgabe ○ keine Vorgabe
Vorbereitung	○ spontan ● vorformuliert ○ vorbereitet ○ vorgelesen
Empraxis	Textvorlage
Feldzugänglichkeit	○ leicht zugänglich ○ frei ● schwer zugänglich ○ unzugänglich

Grafik 1: Beispiel für ein Parameterset in FOLK

Nicht jedes dieser Merkmale ist zur Charakterisierung jeder Gattung relevant. So ist z.B. der Parameter ‚Zeit' für Schlichtungsgespräche irrelevant, und der Parameter ‚zeitliche Ausdehnung' kann nicht a priori fixiert werden. Viele dieser Merkmale sind auch für die Metadatenbeschreibung relevant. Allerdings ist zu bedenken, dass diese Merkmale durchaus nicht für eine Datenaufnahme bzw. eine kommunikative Situation konstant sein müssen (z.B. Gottesdienst: monologische Passagen des Pfarrers vs. Beteiligung der Gemeinde; z.B. Wahlsendung im TV: Moderationen, Interviews, Diskussionen). Hier stellt sich das sehr schwer zu lösende Problem, nach welchen Kriterien Daten segmentiert und als Einheit in FOLK eingestellt werden. Jedes Sprechereignis kann (allerdings nicht exhaustiv!) beschrieben werden durch das Wertemuster, das ihm in der Gesamtmatrix der Parameter zukommt.

Nach dem Prinzip der systematischen Variation solcher Parameter wurden von 1960 – 1977 die im IDS archivierten, unter Leitung von Hugo Steger erhobenen Korpora „Grundstrukturen: Freiburger Korpus" (FR, http://agd.ids-mannheim.de/html/korpora/korpus-fr.shtml) und „Dialogstrukturen" (DS, http://agd.ids-mannheim.de/ html/korpora/ korpus-ds.shtml) zusammengestellt. Die parametrisierte Systematik ist eine etische Systematik. Suche und Zuordnung von Gesprächereignissen geschieht nach Maßgabe von forscherseitig definierten Kategorien. Die Kategorien selbst sollen diejenigen Parameter reflektieren, die für die Konstitution variierender Interaktionsformen entscheidend sind. Ihre Validität bemisst sich also daran, ob sie konstitutionstheoretisch gesehen relevant sind, und ob es gelingt, die konstitutionstheoretisch relevanten Faktoren mit ihnen einigermaßen vollständig zu erfassen. Es handelt sich bei ihnen um weitgehend „unabhängige" Merkmale, d. h., solche, die weitgehend apriorisch, aus der Kenntnis der Situation verifiziert bzw. abgeschätzt werden können und nur eine recht oberflächliche Analyse des in der Situation stattfindenden Kommunikationsereignisses verlangen. Voraussetzung ist dagegen oftmals ein breites ethnographisches oder kulturelles Wissen. Diese etische Systematik geht von einem statischen Situations- und Kontextmodell aus, d. h., die relevanten Merkmale gelten unabhängig vom konkreten Gesprächsprozess als stabil und sollen diesen determinieren.

Eine *Gattungssystematik* geht dagegen von den kommunikativen Gattungen, nach denen Kommunikationsereignisse organisiert sind, aus. Dieser Ansatz entspricht LUCKMANNS (1986) Vorstellung des kommunikativen Haushalts.[1] Gattungen sind emische Orientierungska-

1 Im Rahmen dieses Artikels wird „Kommunikative Gattung" als gesprächstypologischer Sammelbegriff benutzt. Wir differenzieren hier nicht zwischen ‚Gattungen', ‚Kommunikationssituationen', ‚Sprech- bzw. Kommunikationsereignissen', ‚Hand-

tegorien, die vielfach auch mit einer Ethnokategorie bezeichnet werden. Über Parameter der Außenstruktur sind sie mit Merkmalen von Sprechern und anderen sozialstrukturellen Gegebenheiten (wie Institutionen und Milieus) verknüpft, über Parameter der situativen Realisierungsebene mit Parametern wie den o.g. Situationsparametern verbunden (GÜNTHNER/KNOBLAUCH 1994). Die Gattungssystematik entsteht aber nicht über eine systematische Variation dieser Parameter. Die Spezifik der Gattungen beruht nämlich zum einen mindestens ebenso sehr auf ihrer Binnenstruktur, d. h. auf den sprachlich-kommunikativen Verfahren ihres Vollzugs. Zum anderen wird mit dem Gattungskonzept keineswegs impliziert, dass eine systematische Variation von situativen bzw. außenstrukturellen Parametern dazu führt, die gesellschaftliche Kommunikationspraxis abbilden zu können: Einerseits kommen viele Kombinationen gar nicht vor, und andererseits sind viele, z.B. mit Gesprächsthemen, institutionellen Handlungsaufgaben oder besonderen Beziehungsmustern der Beteiligten verbundenen Variationen des Kommunizierens überaus wichtig, die nicht in einer universalen Parametermatrix erfasst werden können. Wesentliches Kriterium für die Differenzierung von Gattungen ist ihre Funktion bzw. ihr Zweck; allerdings sind auch thematische, formal-verfahrensbezogene und mediale Kriterien überaus relevant. Trotz situativer und außenstruktureller Restriktionen von Gattungen sind viele Gattungen dynamisch und emergent, d. h., ihr Auftreten ist nicht oder bestenfalls als Möglichkeit aufgrund vorgängiger situativer Konstellationen antizipierbar. Z.B. sind ‚Streit', ‚Klatsch' und ‚gemeinsame Medienrekonstruktion' Gattungen, die unter bestimmten Situationsparametern entstehen können, aber keineswegs müssen. Außerdem sind Gattungen in der Regel kleinere Einheiten als Sprechereignisse: Sprechereignisse (wie z.B. ein Arzt-Patient-Gespräch) bestehen zumeist aus mehreren Gattungen (z.B. Anamnese, Befunderhebung, Diagnosemitteilung, Verordnung/Therapieplanung, Small Talk). Gerade aus taxonomischer Sicht besteht das Problem, dass Teile einer Gattung in der Literatur auch selbst als Gattung bezeichnet werden (z.B. Vorwürfe Є Streitgespräch) und dass die Zuordnung dabei oft auch noch multipel sein kann (z.B. Sprichwort Є Erzählung und Sprichwort Є Argumentation).

lungsschemata', ‚big packages', ‚activity types' oder ‚Interaktionstypen', obwohl uns natürlich bewusst ist, dass diese Begriffe jeweils unterschiedliche Gegenstandsaspekte akzentuieren und teilweise zu unterschiedlichen gesprächstypologischen Kategorien gelangen, was z.B. die Bindung an außersprachliche Kontexte und Sozialstruktur oder die Größenordnung von Gesprächsphänomenen angeht (s. DEPPERMANN/SPRANZ-FOGASY 2001 für einen Vergleich der Konzepte).

Für die Stratifikation eines ausgewogenen Korpus ist es notwendig, Datentypen taxonomisch zu differenzieren und hierarchisch zu kategorisieren. Dafür gibt es mehrere Gründe:

- Es ist unmöglich, für jede kommunikative Gattung Beispiele zu erheben. Um die vorhandenen Ressourcen gezielt einsetzen zu können, müssen die wesentlichen Distinktionen der oberen hierarchischen Ebenen einer Taxonomie des kommunikativen Haushalts bekannt sein. Es muss versucht werden, Token solcher Gattungen zu erheben, mit denen eine Variation hinsichtlich der Oberkategorien der Gattungstaxonomie im Korpus repräsentiert wird (z.B. dyadische vs. Mehrparteien-Gespräche unter vertrauten vs. unvertrauten Personen etc.). Dieses Kriterium kann ergänzt werden durch das Ziel, Gattungen zu erfassen, die gesellschaftlich besonders relevant sind oder besonders häufig stattfinden.
- Die taxonomische Ordnung ist auch für die Architektur des Metadatenschemas wichtig. Um die Stratifikationssystematik analytisch nutzen zu können (z.B. für gattungsbezogene oder parametergesteuerte Vergleiche), muss sie möglichst explizit sein. Dann kann sie in die Metadatendeskription des Korpus eingehen.

Aus wissenschaftstheoretischen Gründen sollte eine Taxonomie (u. a.) folgenden Kriterien genügen (vgl. ISENBERG 1983 für die ersten drei; GÜNTHNER/KNOBLAUCH 1994 für das vierte):

- *Monotypisierung*: Jedes Datum wird eineindeutig einer taxonomischen Kategorie zugeordnet; Mehrfachkategorisierungen sind ausgeschlossen.
- *Exhaustivität*: Die Taxonomie erfasst alle zu kategorisierenden Daten.
- *Homogenität*: Die Taxonomie ist hierarchisch monoton, d. h. die hierarchischen Kriterien sind strikt implikativ und transitiv; heterarchische Verhältnisse sind ausgeschlossen.
- *Teilnehmerrelevanz*: Taxonomische Kategorien sollen Größen sein, an denen sich die Gesprächsteilnehmer selbst bei der Herstellung der Gesprächsstruktur orientieren.

Die Erfüllung der ersten drei Kriterien wäre für eine systematische Stratifikation gerade auch aus dokumentatorischen und datenbanktechnischen Gründen sehr wünschenswert. De facto ist aber anzunehmen, dass keines dieser Kriterien zu erfüllen ist. Dies zeigt ein Blick in die Textlinguistik, die seit den 60er Jahren Textsortentaxonomien entwickelt und dabei immer wieder festgestellt hat, dass vor allem der dritte Punkt ‚Homogenität' nicht zu erreichen ist, weil Textsorten sowohl

nach Funktionstypen als auch nach Situationstypen, Verfahrenstypen, medialen Typen und thematischen Typen geordnet werden können (vgl. HEINEMANN/VIEHWEGER 1991). Eine Typenkategorie kann aber dabei nicht konsistent als Unterkategorie der anderen modelliert werden, da die Dimensionen nicht orthogonal zueinander sind (z.B. ist nicht jedes Thema unter jeder medialer Bedingung und bzgl. jeder kommunikativen Funktion möglich). Zudem führt das Kriterium ‚Exhaustivität' dazu, dass viele idiosynkratische und Ad hoc-Kategorien gebildet werden müssen, die die Homogenität der Taxonomie einschränken. Dies ist umso ernster zu nehmen, als eine ethnomethodologisch brauchbare Gattungstaxonomie das vierte, emische Kriterium der Teilnehmerrelevanz erfüllen muss. Dies führt unweigerlich dazu, dass gewisse abstraktionshierarchische Kategorien „leer bleiben" und dass andererseits von der Taxonomie her unsystematisch erscheinende Distinktionen eingeführt werden, die von der spezifischen soziokommunikativen Differenzierung eines bestimmten Handlungsfeldes herrühren. Z.B. beruhen die Gattungen ‚Witzerzählung' (SACKS 1978) und ‚Fiktionalisierung' (KOTTHOFF 1998) zentral auf dem Merkmal ‚unernste Modalität', welches wohl für bestimmte Arten von Erzählungen relevant ist, für fast alle anderen Gattungen aber keine relevante Subdifferenzierung darstellt. Ebenso ist die Differenzierung von Witzen und Fiktionalisierungen aufgrund von Merkmalen wie Pointenkonstruktion oder Extendierbarkeit sehr spezifisch für diese beiden Gattungen.

Zu Beginn dieses Abschnitts wurden der parametrisierte und der Gattungsansatz als zwei alternative Modelle eingeführt. In welchem Verhältnis steht das Vorhaben der Taxonomiebildung zu ihnen? Bei genauerer Betrachtung stellt sich heraus, dass sich durchaus beide Systematiken innerhalb einer Taxonomie aufeinander abbilden lassen, wenn man sie entsprechend interpretiert. In Bezug auf das Parametermodell erscheint es dabei v.a. wichtig, das statische Kontextmodell zugunsten eines dynamischen zu verabschieden (vgl. DEPPERMANN/SPRANZ-FOGASY 2001) und die Relevanz emischer Distinktionen zwischen Gattungen ernst zu nehmen, welche im Parametermodell aufgrund ihrer Nicht-Universalisierbarkeit und Irregularität unter den Tisch zu fallen drohen. Viele Werte auf den Parameter-Kontinua lassen sich aber als hierarchisch hohe bis höchste Kategorien einer Gattungstaxonomie ansetzen, so z.B. die Distinktionen ‚privat' – ‚institutionell' – ‚öffentlich'. Gattungen selbst können zu Teilen, aber sicher nicht exhaustiv als Kombinationen von Merkmalsausprägungen auf den Parameter-Dimensionen beschrieben werden, wobei stets einige Merkmale indifferent sein werden, andererseits aber weitere idiosynkratische Merkmale zur Gattungsdefinition hinzukommen müss-

ten (s. u.). Die Bildung der Taxonomie sollte dabei konzertiert von beiden Ansätzen her erfolgen:

- aufgrund einer Sammlung relevanter Gattungen, die gemäß konstitutiver Parameter codiert werden,
- aufgrund der systematischen Kombination von Parameterausprägungen, für die dann entsprechende (beispielhafte) Gattungen gesucht werden (sofern existent).[2]

Für die Umsetzung einer Gattungstaxonomie in ein konkretes Stratifikationskonzept ist allerdings zu beachten, dass die Parameter, die aufgrund externer (statischer, erwartbarer) Gegebenheiten der Interaktionssituation vorherzusagen sind (wie z.B. bei den meisten Gattungen institutioneller Interaktion) wichtiger sind als die dynamischen, emergenten Parameter, die sich in Abhängigkeit von unvorhersehbaren Gesprächsverläufen ändern können. Nur erstere können systematisch zur Entscheidung über die Erhebung von Daten herangezogen werden, da es nicht möglich ist, emergente Ereignisse wie Familienstreits oder Fiktionalisierungen geplant zu erheben. Will man aber trotzdem Sorge tragen, dass relevante emergente Gattungen im Korpus repräsentiert sind, dann muss man sich zunächst fragen, unter welchen externen Merkmalen sie am verlässlichsten auftreten, um sie anhand jener mit einer einigermaßen großen Wahrscheinlichkeit im Korpus erfassen zu können (Will man z.B. rituelle Beschimpfungen haben, muss man versuchen, informelle Kommunikation in jugendlichen Peer-Groups zu erfassen. Ersteres ist nicht systematisch zu erheben, letzteres durchaus.)

3.2 Merkmale der Sprecher

Eine zweite, zu den Merkmalen der Sprechereignisse komplementäre Stratifikationsdimension ist die Variation der Merkmale der Sprecher. Ein ausgewogenes Korpus muss auch ausgewogen sein in Hinblick darauf, dass alle Bevölkerungsgruppen in ihm vertreten sein müssen. Dies gilt sowohl für soziodemographische Kategorien als auch für lebensweltlich fundierte Sprachgemeinschaften und soziale Netzwerke. Dabei ist zum einen die Variation klassischer Parameter der sozialen Struktur zu beachten: Alter, Geschlecht, ethnische bzw. nationale Abstammung, Bildungsstand, Beruf und Einkommen. Das Korpus muss die Variation, idealiter auch die Proportion der Ausprägungen dieser Parameter, die unsere Gesellschaft kennzeichnet, repräsentieren (vgl.

2 Zu diesem Zweck haben wir am IDS eine kleine Datenbank angelegt, in der über 100 Gattungen nach den o.g. Parametern beschrieben sind. Wir danken Franziska Emrich für die Mitarbeit an der Datenbank.

die Veröffentlichungen des Statistischen Bundesamts)[3]. Dies ist zum einen in Hinblick auf die soziale Ausgewogenheit des Korpus nötig. Zum anderen ist anzunehmen und in vielen Punkten auch schon sprachwissenschaftlich erwiesen, dass die Variation dieser Parameter unmittelbar kommunikativ relevant ist, da sich mit den sozialstrukturellen Distinktionen auch Unterschiede in den präferierten kommunikativen Anlässen, Funktionen und Themen verbinden und da unterschiedliche Gruppen und Milieus unserer Gesellschaft unterschiedliche Varietäten (z.B. Fach- und Gruppensprachen) und Kommunikationsstile zur Bearbeitung funktionaler und identitärer Belange ausbilden (vgl. KALLMEYER/KEIM 2002). Die neuere soziologische Forschung weist dabei darauf hin, dass soziodemographische Variablen, die für traditionelle Modelle sozialer Schichtzugehörigkeit grundlegend sind, nicht unbedingt lebensstilistische Orientierungen determinieren. Im Zuge einer zunehmenden Entnormierung und gleichzeitigen kulturellen Ausdifferenzierung der Gesellschaft bilden sich Lebensstile auch unabhängig von Schichtzugehörigkeit aus, bzw. innerhalb ein- und derselben Schicht entwickelt sich eine erweiterte Spannweite differentieller stilistischer Optionen (SCHULZE 1992). Die Zugehörigkeit zu lebensstilistisch geprägten Milieus (vgl. das Konzept der „Sinus-Milieus", http://www.sociovision.de/loesungen/sinus-milieus.html) und Szenen (z.B. im Bereich der Jugendkultur: HITZLER et al. 2005; SCHMIDT/NEUMANN-BRAUN 2004) sowie die Präferenz für gewisse Konsum- und Freizeitorientierungen (z.B. in Bezug auf Sport, Musik, Bildungsangebote) ist für eine realistische Repräsentation der gesellschaftlichen Kommunikationspraxis überaus relevant. Milieus, Szenen und Freizeitgruppen stellen unterschiedliche *communities of practice* (ECKERT 2000: 35ff.) dar, die sich wesentlich über unterschiedliche Kommunikationsstile und -anlässe konstituieren und definieren (z.B. Konzertevents, Tierschauen, Tauschbörsen, Bibellese- oder Selbsterfahrungsgruppen).

Neben diesen allgemein soziodemographisch und sozialstrukturell relevanten Kategorien müssen spezifisch sprach- und kommunikationsbezogene Sprechermerkmale berücksichtigt werden. Hierzu zählen die regionale Herkunft, Muttersprache, Dialektkompetenz, Sprachpathologien (Aphasie, Dysarthrie etc.) und andere kommunikationsrelevante Behinderungen (Hör-, Sehbehinderung, Amnesien). Auch wenn sie nicht systematisch erhoben werden (s. Abschn. 2), ist ihre Dokumentation wichtig, um die Typikalität des aufgenommenen Gesprächs und

3 http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Navigation/Statistiken/Bevoelkerung/Bevoelkerung.psml

Faktoren, die auf seinen Verlauf eine moderierende Wirkung ausgeübt haben können, in der Analyse berücksichtigen zu können.

3.3 Merkmale der Sprache der Interaktion

Ein nationales Gesprächskorpus zielt nicht auf eine systematische Erfassung der arealen Variation und des sich zunehmend ausbildenden Dialekt-Standard-Kontinuums ab (vgl. dazu BEREND 2005). Die Erfassung der Breite kommunikativer Gattungen und Anlässe führt aber dazu, dass verschiedene Grade der Standardnähe und -ferne im Korpus vorkommen, die für die unterschiedlichen Gattungen und Anlässen (z.B. aufgrund von Formalität und Vertrautheit der Beteiligten) charakteristisch sind. Diese Variation gehört nicht einfach zu den Sprechermerkmalen, denn die meisten Sprecher verfügen über ein Spektrum differenziell, d. h., gattungs- und adressatenspezifisch einzusetzender Varianten. Der Realität der (post)modernen multikulturellen Gesellschaft, die durch Migration geprägt ist, muss außerdem durch die Aufnahme von mehrsprachigen Daten (Mischsprachen, Code-Switching, Dolmetschinteraktionen) und Migrantenvarietäten Rechnung getragen werden.

Die Merkmale a)–c) ergänzen sich bei der Korpusstratifikation. In theoretischer Hinsicht ist nur eine Kombination aller drei Merkmalsgruppen geeignet, um ein ausgewogenes Korpus zu erstellen (vgl. BURNARD 2001). Die Gattungssystematik soll einen fundierten Überblick über die qualitative Differenzierung der Kommunikationspraxis in den unterschiedlichen sozialen Sektoren unserer Gesellschaft geben. Sie deutet auch auf Kommunikationsereignisse hin, die zwar nicht häufig vorkommen (Kanzlerkandidatenduell) oder nur von kleinen Sprechergruppen produziert werden (TV-Moderationen, Cockpitkommunikation), die aber von zentraler Relevanz für die gesellschaftliche Kommunikationspraxis sind. Naturgemäß gilt dies für vor allem für massenmediale Kommunikationsereignisse. Die Systematik der Gattungen ist weiterhin nötig für eine systematische Erhebung des Bereichs institutioneller Kommunikation.

Andererseits erlaubt nur der Ausgang von den Sprechermerkmalen, die kommunikativen Zeitbudgets und Routinen unterschiedlicher Bevölkerungsgruppen und die soziale, d. h. außenstrukturell bedingte Differenzierung von Gattungen (sowohl hinsichtlich ihrer speziellen Ausprägung als auch hinsichtlich ihrer Praktizierung überhaupt) zu erfassen. Zu vielen Kommunikationsereignissen ist ja überhaupt nur ein Zugang zu gewinnen, wenn sie am Leitfaden des Handelns eines Sprechers, d. h., als aufeinander folgende Momente seines Tageslaufs er-

fasst werden. Gerade die lebensweltlich relevanten Distinktionen der kommunikativen Praxis sind fast nur ausgehend von Sprechern, die in den betreffenden Milieus verankert sind, zu erheben, nicht aber durch einen „direkten" Zugang zum Kommunikationsereignis als solchem, da dieses nicht vorhersehbar oder für Außenstehende gar nicht zugänglich ist.

Die Variation der Sprechervariablen ist genauso wie die Variation der Gattungen bzw. der Parameter der Kommunikationssituation auf eine qualitative Ausgewogenheit des Korpus angelegt. Das erste und bereits überaus ehrgeizige Ziel der Korpusstratifikation ist es, die unterschiedlichen Sprechergruppen überhaupt erst einmal als solche zu repräsentieren, d. h., eine möglichst große Vielfalt im Korpus aufzunehmen. Erst zu einem späteren Zeitpunkt kann man darüber nachdenken, wie man statistisch relevante Stratifikationsparameter gewinnt. Dazu müssten makrosoziologische Parameter berücksichtigt werden, wie z.B. gesellschaftliche Zeitbudgets (bestimmter sozialer Gruppen für bestimmte kommunikative Aktivitäten), soziodemographische Verteilungen, die sektorielle Logik der gesellschaftlichen Praxis usw. (s. aber Abschn. 5).

4. Stratifikationskonzepte bestehender Korpora

Bevor wir die Probleme und Möglichkeiten der in Abschn. 3 vorgestellten Zielperspektive diskutieren (s. Abschn. 5–7), wollen wir zunächst einen Blick auf die Vorgehensweise der Stratifikation bzw. die Bestände von vorhandenen großen nationalen Gesprächskorpora werfen. Dadurch wird deutlich, welche Konzepte bereits entwickelt und umgesetzt wurden, ebenso aber auch, an welche Grenzen man dabei stieß. Im Folgenden werden drei Korpora diskutiert: BNC (Großbritannien), CGN (Niederlande) und GSLC (Schweden). Im Unterschied zu anderen großen Korpora sind diese nämlich unserem Vorhaben darin vergleichbar, dass

- sie nach einem systematischen Stratifikationsprinzip zusammengestellt werden und nicht (nur) eine Sammlung verfügbarer Korpora anbieten,
- sie Ausgewogenheits- und Repräsentativitätskriterien zu berücksichtigen versuchen sowie
- Gesprächsdaten einen wesentlichen Bestand des Korpus bilden und sich die Stratifikationsüberlegen vor allem auf diesen Datentyp beziehen.

4.1 British National Corpus (BNC)

Das BNC hat das bisher avancierteste Stratifikationskonzept verfolgt (http://www.natcorp.ox.ac.uk/). Ca. 10% seines Gesamtumfangs von 100 Mio. Wörtern sind gesprochensprachlich. Korpusdesign, technische und Transkriptionsstandards etc. sind auf den Webseiten des BNC ausführlich beschrieben. Das mündliche Korpus mit einem Zielumfang von 10 Mio. Wörtern wurde nach zwei Kriterien zusammengestellt, nach denen jeweils etwa die Hälfte des Bestands erhoben wurden:[4] Nach „text types" (= Gesprächstypen) und gemäß der soziodemographischen Merkmale Alter, Geschlecht, soziale Schicht und Region innerhalb Großbritanniens.

Für das *soziodemographische Sampling* wurden 153 Informanten (auch unter 16 Jahren) ausgewählt, die gebeten wurden, eine Woche lang alle Interaktionen, an denen sie beteiligt sind, aufzunehmen. Die Gesprächspartner sollten jeweils im Anschluss ans Gespräch um Einwilligung gebeten werden.[5] Die Informanten mussten ein Tagebuch führen, in dem Teilnehmer, Zeit und Setting der Aufnahmen verzeichnet wurden. So wurden insgesamt 700 Stunden (3.56 Mio. Wörter) aufgenommen. Insgesamt sind über 1000 Sprecher auf den Aufnahmen. Der Aufnahme ging eine Pilotstudie voraus, in der Fragen der Rekrutierung, der Datenaufnahme und Protokollierung sowie der Dokumentation optimiert und erfahrungsbasierte Schätzwerte für den Aufwand der Erhebung und Aufbereitung des Gesamtkorpus gewonnen wurden.

Der zweite Teil des Korpus wurde nach einer *Taxonomie von text types* erhoben („context-governed part of the corpus"). Die vier hierarchisch höchsten Kategorien sind *educational, business, public/institutional, leisure*. Dabei wurden je 40% monologische und 60% dialogische Daten angestrebt. Die Daten sollten nach den Kriterien Region, Geschlecht, Bildungsniveau der Sprecher sowie Gesprächsthema balanciert werden. Die einzelnen Aufnahmen sollten nicht mehr als 10000 Wörter beinhalten, pro *text type* wurden 200000 –300000 Wörter angestrebt. Insgesamt wurden so 757 Aufnahmen mit 6.15 Mio. Wörtern für das BNC gewonnen. Im Einzelnen umfasst der BNC folgende *text types*:

- *educational: lectures/talks/demonstrations, news commentary, classroom interaction*
- *business: company talks and interviews, trade union talks, sales demonstrations, meetings, consultations*

4 s. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#spodes

5 Dies widerspricht deutschen Datenschutzbestimmungen, nach denen immer *vor* der Aufnahme die Einwilligung eingeholt werden muss.

- *public/institutional: political speech, sermons, public/ government talks, religious meetings, parliamentary and legal proceedings*
- *leisure: speeches, broadcast sports commentaries, talks to clubs, phone-ins, broadcast chat shows, club meetings.*

4.2 Corpus Gesproken Nederlands (CGN)

Das niederländische Nationalkorpus CGN wurde zwischen 1998 und 2004 für 4.6 Mio € erstellt (im Überblick: OOSTDIJK 2002). Es umfasst 8.916 Mio. Wörter. Das gesamte Korpus ist text-ton-aligniert. Die Korpusstratifikation richtet sich nach dem *socio-situational setting* (http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm#intro) , das nach den Parametern Gesprächsziel, Medium, Anzahl der Teilnehmer und Sprecher-Hörer-Beziehung beschrieben wird. Das Ziel bestand dabei weniger darin, ein statistisch ausgewogenes Korpus zu erstellen als vielmehr eine breite Varianz an Daten zu gewinnen. Heterogenere Datentypen sind breiter repräsentiert. Die Sprechermerkmale Geschlecht, Alter, Region, Schicht und Bildungsabschluss sollen berücksichtigt worden sein bei der Korpuszusammensetzung, es wird aber nicht dargelegt, wie. Im Korpus enthalten sind folgende Daten:

- *spontaneous face-to-face conversation*
- *spontaneous telephone conversations*
- *interviews with teachers*
- *simulated business negotiations*
- *broadcast interviews/discussions/debates*
- *political discussions/debates/meetings*
- *classroom interaction*
- *broadcast live commentaries*
- *broadcast news/reports/reportages*
- *broadcast commentaries/reviews/columns*
- *ceremonious speech/sermons*
- *university lectures/seminars*
- *read speech* (Vorlesesprache)

Mehr als die Hälfte des Korpusumfangs (4.7 Mio. Wörter) besteht aus den ersten beiden Kategorien spontaner Alltagskonversation. Die Spannweite der Daten ist grundsätzlich ähnlich wie beim BNC, allerdings fehlen einige Gattungen und es sind Rollenspieldaten enthalten.

4.3 Göteborg Spoken Language Corpus (GSLC)

Das GSLC umfasst 1.42 Mio. Wörter (182h) von erwachsenen Sprechern mit Schwedisch als Muttersprache. Zusätzlich gibt es noch kleinere Sonderkorpora (Schwedisch als L2, interkulturelle Kommunikation, Kindersprache, pathologische Daten (Aphasie/Dyslexie/Hörbehinderte)). Das Korpus wird opportunistisch wachsend, auf der Basis v.a. von Aufnahmen aus Forschungsprojekten und studentischen Arbeiten, ohne ein übergreifendes Stratifikationskonzept mit dem Ziel zusammengestellt, sukzessive noch nicht enthaltene Gattungen aufzunehmen. Ziel ist eine möglichst breite Spannweite unterschiedlicher sozialer Aktivitäten. Zu Kodierung, Tagging, Transkription, des Browsers und der Auswertung des Korpus gibt es eine Reihe an Publikationen (siehe http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=5). Eine Zusammenfassung der Transkripions- und Codierungsstandards und -tools des Korpus findet sich in ALLWOOD et al. (2000). Folgende Aktivitätstypen sind im Korpus enthalten:

- *Discussion: broadcast, tv, university seminar, students' focus group*
- *Retelling of Article*
- *Interview: biographical, thematic, broadcast*
- *Task-Oriented Dialogue: experiments: object specification, airplane*
- *Informal Conversation: school bus, quarrel, family, marital, among friends, etc.*
- *Role Play*
- *Trade Fair*
- *Arranged Discussions: mass media*
- *Formal Meeting: administrative, company, district committee, fund board, government agency, municipality, local authority*
- *Consultation: doctor-patient (different disciplines), optician*
- *Shop: food store, game shop, radio/tv shop, supermarket*
- *Dinner*
- *Flea Market*
- *Auction*
- *Factory Conversation*
- *Party*
- *Games & Play: computer game*
- *Phone: Alarm, car hire, insurance, restaurant, private*
- *Travel Agency: phone, face-to-face*
- *Court: trial*
- *Church: sermon*

- *Lecture*
- *Hotel: kitchen, reception, conference department*
- *Physiotherapy, occupational therapy*
- *Bus Driver-Passenger*
- *Reality tv*

4.4 Evaluation der Stratifikationskonzepte

Das expliziteste Konzept und den überzeugendsten Datenbestand bietet das BNC. Die zweigleisige Stratifikationsstrategie (Ausgehend vom Genre und von den Sprechern) bietet einen diversifizierten Zugriff auf unterschiedliche Ausschnitte der Gesprächwirklichkeit, vor allem auch auf die nicht systematisch zu erhebenden Gattungen. Bei der Darstellung des CGN fällt eine sehr unzulängliche Gattungsdifferenzierung auf, beim GSLC wird eine Liste präsentiert, die aus einer Mischung von Aktivitäts- und Settingparametern besteht. Dabei scheint die Datenspannweite des GSLC attraktiver zu sein, da dieses Korpus eine größere Vielfalt institutioneller Interaktionsformen anbietet. Bei beiden Korpora ist nicht erkennbar, auf welcher theoretisch-systematischen Grundlage die Liste der erhobenen bzw. zu erhebenden Daten erstellt wurde. Es ist auch nicht erkennbar, nach welcher interaktionstypologischen Systematik die Daten im Korpus geordnet sind.

5. Theoretische Probleme des Korpusdesigns

5.1 Repräsentativitätsproblem

In der Korpuslinguistik besteht breiter Konsens, dass Repräsentativität im statistischen Verständnis kein Ziel einer Korpuskonstruktion sein kann (vgl. LEMNITZER/ZINSMEISTER 2006). In Bezug auf ein nationales Gesprächskorpus ist erstens unklar, wie die zugrunde liegende Grundgesamtheit, aus der eine Stichprobe gezogen werden soll, zu bestimmen ist:

- Es gibt keine Liste aller Gattungen der kommunikativen Praxis. Das diesbezügliche Wissen ist überaus verstreut und großenteils intuitiv, nirgends niedergelegt. Zwar ist es für eine systematische Stratifikation zwingend erforderlich, einen Überblick über den gesellschaftlichen Bestand kommunikativer Gattungen zu gewinnen, doch sollte dieses Ziel nicht verdecken, dass a) Gattungsbezeichnungen und -grenzen vielfach perspektivisch und umstritten sind und dass in der kommunikativen Praxis unaufhörlich neue Gattungen entstehen (z.B. Hiphop-

Battles, Doku-soap-Formate) und alte transformiert werden (z.B. Fernsehbeichten) und vergehen (z.B. bürgerliche Antrittsbesuche, Duellforderungen).
- Es gibt keine Abschätzungen des quantitativen Vorkommens einzelner Gattungen.
- Die quantitative Zusammensetzung der Gesprächswirklichkeit kann bzgl. vieler Parameter kaum sinnvoll abgeschätzt werden. (Wie ist z.B. die Alters- und Geschlechtsverteilung der älteren Bezugspersonen beim Bilderbuchvorlesen? Wie viele Autofahrer schimpfen wie oft beim Fahren, wie viele singen dabei wie oft?)

Zweitens ist unklar, nach welchen Kriterien verschiedene Gesprächsgattungen zueinander im Korpus gewichtet werden sollen. Wie sollen z.B. medial distribuierte vs. *face-to-face*-Gattungen, wie Gespräche mit zwei vs. mehreren Teilnehmern zueinander gewichtet werden? Wie sind Häufigkeit, Teilnehmerzahl, aktive Produktion vs. reine Rezeption, das Verhältnis zu evtl. simultanen Aktivitäten etc. zueinander zu gewichten? Massenmediale Produkte werden beispielsweise nur von einer sehr kleinen Sprecherzahl produziert, dafür kann die Zahl der Rezipienten in die Millionen gehen.

Drittens ist es vollkommen illusorisch, eine vollständige Merkmals-Kombinatorik im Sinne einer Kreuztabellierung aller Gattungen mit allen Sprechermerkmalen (soweit empirisch vorkommend) anzustreben. So ist es z.B. unmöglich, jede Gesprächsgattung in allen arealen Variationen, mit allen möglichen Kombinationen von Alters- und Geschlechtszusammensetzungen der Sprecher zu erheben.

Eine wesentliche Aufgabe der theoretischen Fundierung von FOLK ist es daher, ein Konzept von ‚Ausgewogenheit' bzw. ‚qualitativer Repräsentativität' im Sinne des LUCKMANNSCHEN Konzepts des ‚kommunikativen Haushalts' zu entwickeln, das keine falsche Repräsentativitätssuggestion vorgaukelt, aber dennoch eine theoretisch wie empirisch fundierte Idee vermittelt, welchen Anspruch der Abbildung kommunikativer Wirklichkeit des Deutschen FOLK stellen will und warum dieser Anspruch relevant ist. Das vorliegende Papier ist ein erster Versuch dazu.

5.2 Probleme des konstitutionstheoretischen Zugriffs

Das Problem der Grundgesamtheit ist nicht nur ein empirisch-quantitatives Problem. Die entsprechenden Fragen des Korpusdesigns berühren auch konstitutionstheoretische Probleme. Der taxonomische Ansatz präsupponiert, dass kommunikative Gattungen die relevante

Ebene sind, auf der die kommunikative Praxis vollständig zu erfassen und zu ordnen ist. Das ist aber durchaus fragwürdig, selbst wenn man, wie wir es hier tun, von einem sehr weit reichenden Gattungsbegriff ausgeht (vgl. Fußnote 1). Es gibt nämlich viele kommunikative Episoden, die nicht gattungsmäßig organisiert sind. Das betrifft vor allem Transitions-, Anbahnungs- und Nachphasen von Gesprächen, die nicht klar einer Gattung zuzuordnen sind, selbst aber keinen eigenständigen Ereignisstatus haben (vgl. MONDADA/SCHMITT 2010). Es betrifft weiterhin kommunikative Episoden, für die nicht zu erwarten ist, dass sie sich an Gattungserwartungen orientieren (wie z.B. ein Gespräch von Fahrradfahrern während des Wartens an der Ampel), und kommunikative Kurzereignisse (wie eine Entschuldigung in der Straßenbahn, selbstbezogene Bemerkungen beim Autofahren etc.). Diese Ereignisse sind hinsichtlich ihres kategorialen Status sehr unklar (teilweise keine fokussierten Interaktionen, keine klaren Grenzen), kennen meist keine Gesprächstypbezeichnung oder werden anhand von nicht-verbalen Situationsparametern bezeichnet („Gespräche beim Umkleiden", „in der Warteschlange"...). Sie sind kaum untersucht und oft nicht systematisch zu erheben, scheinen aber einen bedeutenden Teil der alltagsweltlichen Kommunikationspraxis auszumachen.

Ein weiteres Problem ist, dass empirische Kommunikationsereignisse oft hochgradig hybrid sind: Wechsel zwischen aufgabenbezogener und phatischer Kommunikation, zwischen Narration und Argumentation, empraktischen Instruktionen und Smalltalk etc. sind an der Tagesordnung. Oft ist daher weder eine eindeutige kategoriale Zuordnung solcher „Mischformen" möglich, noch ist anhand scheinbar statischer Situationsparameter gesichert, dass dann im Gespräch auch wirklich (nur) passiert, was durch die Situationsparameter von Thema, sozialen Rollen und Gesprächszweck eindeutig projiziert zu sein scheint (z.B. Smalltalk im Arzt-Patient-Gespräch, Witze erzählen in Besprechungen etc.; vgl. SCHEGLOFF 1991). Die Erkenntnis, dass Kontexte reflexiv und dynamisch hergestellt und somit oft auch modifiziert werden (vgl. AUER 1992), erschwert die Möglichkeit einer systematisch nach apriorischen Kriterien verfahrenden Stratifikation immens, da in vielen Fällen eben erst dann, wenn man die Aufnahme erhoben hat und kennt, klar ist, *wovon* es eine Aufnahme ist. Diese Flexibilität führt zu einer weiteren Komplexierung, denn es handelt sich nicht einfach nur um kontextuelle Modifikationen. Oft bestimmt vielmehr gerade das Verhältnis der Abweichung gegenüber einem vorgängigen bzw. erwarteten oder funktionalen Kontext die besondere Gestalt und Funktion einer emergenten Form (z.B. ist ein Smalltalk in einer Geschäftsverhandlung nicht nur „unerwartet", sondern vielleicht auch in spezifischer Weise

funktional an den offiziellen Gesprächszweck angepasst und deshalb vielleicht auch formal anders als ein Smalltalk auf einer Stehparty).

5.3 Probleme der Parametrisierung

Viele der Parameter der Gesprächssituation, aber auch manche Sprecherparameter bereiten für ein systematisches Korpusdesign (und auch für eine Metadatendeskription) erhebliche Probleme:

- Interpretative Kategorien sind nur schwer zu standardisieren und oft sehr auswertungsintensiv. D. h., sie sind nicht aus externen Merkmalen abzuleiten, sondern würden eine genaue Analyse der erhobenen Daten erfordern. Dies betrifft vor allem Gattungskategorisierungen, aber auch bspw. den Teilnehmerstatus oder die genaue Rollenbezeichnung. Dies ist oft weniger noch ein Problem beim Korpusdesign als bei der Metadatenkategorisierung. Um dieses Problem einzudämmen und die Gattungsstratifikation auf eine intersubjektiv verlässliche und zugleich wenig auswertungsabhängige Basis zu stellen, wird es für das Korpusdesign (und die Metadatenbeschreibung) sinnvoller sein, sich an (eher gröberen) externen Situationsparametern (z.B. „Arzt-Patient-Gespräch", „Unterrichtsgespräch") zu orientieren als an (feineren, im Gesprächsverlauf häufiger wechselnden) Parametern des Gesprächsprozesses („Anamnesegespräch", „fragend-entwickelnder Unterricht").
- Die Ausprägung mancher Parameter (insbesondere bzgl. relevanter Sprechereigenschaften wie lebensstilistische Zugehörigkeiten, wechselseitige Kenntnis) und ihre Relevanz für ein bestimmtes Gesprächsereignis kann a priori nicht abgeschätzt werden.
- Viele Datentoken müssten aus empirischen Gründen aufgrund ihrer inneren Mischstrukturiertheit und aus theoretischen Gründen aufgrund der Heterarchie der Taxonomie mehrfach kategorisiert werden. Dies ist v.a. für die Metadatendeskription hochproblematisch, da eine liberale, Mehrfachcodierungen erlaubende Kategorisierung gewaltige Probleme der Recherchierbarkeit und Auswertbarkeit des Korpus nach sich zieht (uneindeutige Resultate, irreführende Zuordnungen von Datensegmenten zu Metadaten, scheinbare Multiplikation von Tokens etc.).

6. Praktische und forschungsethische Probleme der Korpusstratifikation

Selbst wenn wir die in 5. dargelegten theoretischen Probleme der Korpusstratifikation gelöst hätten, ergäbe sich aus einer theoretisch begründeten Stratifikationssystematik noch lange kein umsetzbares Konzept des Korpusaufbaus. Dem stehen praktische und forschungsethische Hindernisse entgegen, die dazu führen, dass wohl nur ein Bruchteil (!) der gesellschaftlichen Kommunikationspraxis überhaupt bzw. mit dem für ein großes Korpuserstellungsprojekt (das ja selbst kein Gesprächs*forschungs*projekt ist!) vertretbaren Aufwand erhoben werden kann.

6.1 Rechtliche und ethische Anforderungen

Die Erstellung eines öffentlichen, via Internet verfügbaren Korpus stellt besondere Anforderungen an die rechtliche Autorisierung der Datenveröffentlichung. Gleichzeitig radikalisiert diese Form der Verfügbarmachung von Primärdaten ethische und rechtliche Probleme, die im Kern schon von der Veröffentlichung von Textbänden und Transkripten in Büchern gelten, da Missbrauchsmöglichkeiten und potenzielle Gefährdungen für die Aufgenommenen ungleich größer als bei traditionellen Publikationsformen sind. Die öffentliche Nutzung von Gesprächsdaten ist gemäß *Datenschutzrecht* nur dann rechtlich einwandfrei autorisiert, wenn *informed consent* vorliegt: Bereits vor der Aufnahme muss von allen Aufgenommenen (nach Möglichkeit in schriftlicher Form) eine Einwilligung eingeholt werden, in der die Verwendungszwecke der Aufnahme (Nutzung für wissenschaftliche Zwecke) und die mediale Form der zur Veröffentlichung autorisierten Daten (via Internet, Transkript, Audio und/oder Video) festgelegt werden. Es liegt auf der Hand, dass entsprechende Einwilligungserklärungen vielfach gar nicht bzw. nicht von allen Aufzunehmenden zu erlangen sind. Für viele Interaktionssituationen, insbesondere solche mit vertraulichem, für die beteiligten Personen oder Institutionen/Organisationen folgenreichen Inhalten (z.B. Gerichts-, Tarif- oder Geschäftsverhandlungen, psychotherapeutische und medizinische Gespräche, Paarinteraktionen) ist keine Einwilligung zur Datennutzung für eine anonyme Wissenschaftsöffentlichkeit zu erhalten, sondern allenfalls die personalisierte Erlaubnis zur Nutzung innerhalb eines Forschungsprojekts. Selbst wenn eine Einwilligung vorliegt, ist außerdem zu beachten, dass Korpusinhaber aus forschungsethischen Gründen darauf achten sollten, die Aufgenommenen vor potenziellem Schaden

zu bewahren, so dass in manchen Fällen selbst autorisierte Daten nicht zu nutzen sind (z.B. wenn von Straftaten oder psychischen Problemen berichtet wird). Insbesondere bei Aufnahmen von öffentlichen Interaktionssituationen mit anonymen und fluktuierenden (und deshalb nicht a priori vorhersehbaren) Beteiligungsstrukturen ist es gar unmöglich, alle potenziell Aufzunehmenden auf eine Einwilligung anzusprechen. Weitere Einschränkungen entstehen daraus, dass manche Daten praktisch nicht zu anonymisieren sind, sei es aus inhaltlichen oder medialen Gründen: Manche Gattungen sind derart stark von anonymisierungsbedürftigen Inhalten geprägt, dass deren Tilgung zur Unverständlichkeit des Datums führen würde (wie biographische Interviews, Reklamationsgespräche oder Produktplanungsbesprechungen); Videoaufnahmen mit unkenntlich gemachten Gesichtern sind für die meisten Zwecke der Analyse multimodaler Interaktion unbrauchbar. Die wachsende Skepsis gegenüber Datenmissbrauch in Öffentlichkeit und Institutionen macht es zudem zunehmend schwerer und aufwändiger, Einwilligungen gerade auch von institutionellen Akteuren zu erhalten. Aus datenschutzrechtlichen Gründen ist es auch nicht möglich, bereits vorhandene, aber für einen anderen Zweck erhobene Daten weiter zu verwenden, da solche Daten nicht ausdrücklich für die Verfügbarmachung in einem öffentlichen Korpus autorisiert sind (und oftmals bei vorhandenen Korpora gar keine (schriftlichen) Einwilligungserklärungen vorliegen).

Gespräche, die über die Massenmedien verbreitet werden (Phone-In-Sendungen, Talk-Shows oder Diskussionssendungen), unterliegen dagegen dem *Urheberrecht*. Insbesondere wenn nicht nur kurze Ausschnitte, sondern ganze Sendungen verfügbar gemacht werden sollen, was gerade für gesprächsanalytische, sich auf die Gattung als solche oder einen größeren Interaktionskontext erfordernde Fragestellungen wichtig ist, kann nicht von einem *fair use* ausgegangen werden. Die Inkorporation solcher Daten ins Korpus erfordert Lizenzverträge, die mit erheblichen Kosten verbunden sein können, sofern die Rechteinhaber überhaupt die wissenschaftliche Sekundärnutzung ihrer Sendungen zu autorisieren bereit sind.

6.2. Probleme der Erhebbarkeit von Daten

Nur ein kleiner Teil der Gesprächswirklichkeit kann elizitiert, d. h. durch ForscherInnen selbst veranlasst werden, ohne deshalb an Authentizität einzubüßen. Dazu gehören Interviews, biographische Erzählungen oder Serviceanfragen. Andere Daten können systematisch erhoben werden, weil bestimmte Gesprächsereignisse zu festgelegten bzw. vorhersehbaren Anlässen stattfinden. Dies gilt vor allem für institutio-

nelle Gespräche (Unterrichtsstunden, Arzt-Patient-Interaktionen oder Teambesprechungen). Meistens ist hier allerdings einiges an ethnographischer Vorarbeit notwendig, bis man Aufnahmegenehmigungen bekommt, und es gibt mehr oder weniger große Hindernisse des Datenschutzes. Selbst Typen von Gesprächsereignissen, die häufig stattfinden, sind aber oft nicht systematisch zu erheben. Insbesondere solche Gattungen, die sich emergent im Gesprächsverlauf entwickeln, sind dagegen nicht geplant zu erheben (z.B. Streit- oder Klatschgespräche), sondern können nur mit mehr oder weniger großer Wahrscheinlichkeit gewonnen werden, wenn Aufnahmen in Interaktionssituationen (z.B. Gespräche unter Freunden oder in der Familie) gemacht werden, innerhalb derer erfahrungsgemäß diese Gattungen auftreten. Viele Kommunikationsereignisse und ganze kommunikative Milieus (z.B. Hochadel, Mafia, Führungseliten) sind allerdings für Forscher entweder grundsätzlich unzugänglich oder zumindest nicht für ein öffentliches Korpus zu gewinnen. In anderen Fällen wären die Kosten für den Datengewinn unverhältnismäßig hoch.

6.3 Probleme verfügbarer und repräsentierbarer Datenqualität

Viele Kommunikationsereignisse können nicht in der Datenqualität aufgenommen oder verfügbar gemacht werden, die für eine adäquate Analyse erforderlich ist. Nebengeräusche, Lärm und Parallelgespräche beeinträchtigen die Möglichkeit von Audioaufnahmen für viele Kommunikationsereignisse derart, dass sie nicht zu transkribieren sind bzw. das Tonsignal nur lückenhaft oder schlecht auditiv und gar nicht mit instrumentalphonetischen Verfahren zu analysieren ist. Erst der Blick auf Probleme der zu erzielenden technischen Aufnahmequalität macht deutlich, welch großer Teil unserer Kommunikationspraxis vor einer erheblichen Geräuschkulisse stattfindet (z.B. Gespräche in Kneipen, im Auto oder während der Arbeit, Gespräche bei laufendem Radio oder TV, Gespräche im öffentlichen Raum, z.B. in Diskotheken oder auf Baustellen).

Bei anderen Kommunikationsereignissen besteht das Problem eher im Bereich der Dokumentation. Wenn die Beteiligten nicht für eine ausführliche Befragung zur Verfügung stehen, können Metadaten nur lückenhaft dokumentiert werden, da wesentliche Sprechermerkmale nicht anhand der Aufnahme und ihres Kontexts für den Erhebenden zu erschließen sind. Solche Daten mögen nicht grundsätzlich unbrauchbar sein, ihre spätere Erschließung und Auswertung innerhalb eines Korpus kann aber erheblich eingeschränkt sein.

Manche Kommunikationsereignisse sind dagegen ohne Videoaufnahme oder umfangreiche Zusatzmaterialien nicht sinnvoll zu untersuchen, da die Interaktion wesentlich über visuelle Merkmale organisiert ist, medial vermittelt ist oder auf interaktionsexterne Materialien zurückgreift. Dies ist z.B. der Fall bei medial vermittelten Interaktionen wie Videokonferenzen (MONDADA 2007a) oder fachöffentlich übertragenen Schau-Operationen (MONDADA 2007b), Interaktionen, die die Nutzung von Computerbildschirmen, Grafiken und Texten involvieren (HEATH/VAN LEHN 2008), oder bei handlungsbegleitender Kommunikation, bei der referenzielle Bezüge und Handlungskohärenz nur über das visuelle Geschehen herzustellen sind. Diese Fälle bringen zusätzliche Komplexitäten für den Korpusaufbau mit sich, da für eine angemessene Analyse auch entsprechende Video- und weitere Zusatzmaterialien bereitgestellt werden müssten, was für die Datenverwaltung und den Zugriff zu erheblichen Komplikationen führt. Es wird dann deutlich, dass für die angemessene Untersuchung vieler Interaktionssituationen Zusatzmaterialien und Metadatendokumentationen nicht ausreichen. Vielmehr werden solche Daten nur im Kontext einer ethnographischen Untersuchung auszuwerten sein, da sich der Forscher viel an Kontext- und Hintergrundwissen jenseits der Gesprächsaufnahme aneignen muss, um diese angemessen kontextualisieren zu können (KISSMANN 2009). Ein Beispiel dafür sind etwa Aufnahmen von einem Filmset (SCHMITT 2007).

7. Praktische Prioritäten des Korpusaufbaus

An die Seite von theoretischen, gesprächs- und soziolinguistisch bzw. soziologisch motivierten Kriterien der Korpusstratifikation treten daher pragmatische Restriktionen, die gemeinsam in praktische Entwicklungsprioritäten des Korpusaufbaus umgesetzt werden. Für die Entwicklung von FOLK ergänzen wir theoretische Erwägungen daher um folgende Kriterien:

- Optimierung von Kosten-Nutzen-Relationen,
- rechtliche Unbedenklichkeit,
- möglichst weit reichendes Nutzerinteresse von Daten,
- hochwertige technische Qualität der Datenerhebung.

In der ersten Ausbauphase gilt als Hauptpriorität die möglichst *schnelle Steigerung des quantitativen Bestandes* an Daten. Wir verfahren mit der Datensammlung zunächst opportunistisch, d. h., es werden zunächst leicht erreichbare und kostengünstig zu erhebende Kommunikationsanlässe aufgesucht, die es ermöglichen, möglichst bald einen attraktiven und einigermaßen großen Initialbestand von Daten öffentlich verfügbar

zu machen. Prioritär für die Korpuserstellung ist es dabei zunächst, die grundlegenden Parameter der „Merkmale der Sprechsituation" (s. Abschn. 3.1) möglichst breit zu variieren. Das erste Ziel besteht darin, eine möglichst breite Spannweite unterschiedlicher kommunikativer Anlässe und Formen (Gattungen) im Korpus abzudecken. Hierbei gilt das Prinzip „*Breite vor Tiefe*": Anstelle der systematischen Kartierung eines spezifischen Kommunikationsfeldes (z.B. Arzt-Patient-Gespräche) in all seinen Unterausprägungen (Anamnese-, Therapieplanungs-, Diagnosemitteilungsgespräche in Allgemeinmedizin, Urologie, Zahnmedizin etc.) geht es vielmehr zunächst darum, das Spektrum der Variation der grundlegenden Parameter der Variation von Sprechereignissen grob abzubilden, d. h. also z.B. sowohl private wie auch institutionelle und massenmediale Gespräche, dyadische wie Mehrpersoneninteraktionen, Gespräche unter Fremden wie unter Vertrauten etc. bereit zu halten. Im Jahr 2011 werden wir voraussichtlich 75 Stunden Aufnahmen (vorauss. ca. 500.000 Wörter) als Transkript, mit alignierter Tonaufnahme und versehen mit Metadaten verfügbar machen können. Das Korpus wird sich bis dahin aus privaten Interaktionen im Familien- und Freundeskreis, institutionellen Besprechungen, Unterrichtsgesprächen und Arbeitsinteraktionen zusammensetzen. Für die Konstruktion von Ausgewogenheit innerhalb des Korpus gilt dabei das Prinzip „*Qualitative Repräsentativität vor quantitativer Repräsentativität*": Es wird zunächst nicht angestrebt, Ausschnitte der kommunikativen Wirklichkeit durch eine so hohe Anzahl von Fällen im Korpus zu repräsentieren, dass inferenzstatistisch gestützte Aussagen über die zugrundeliegende Grundgesamtheit gemacht werden können. Dies würde eine sehr hohe Anzahl von Fällen eines sehr spezifischen Ausschnitts der Kommunikationspraxis erfordern, die für viele Gattungen gar nicht oder nur auf Kosten des Spektrums an Sprechereignissen zugunsten einiger weniger quantitativ stark repräsentierter Sprechsituationen zu gewinnen sind.

Da sich das Korpus in erster Linie an Gesprächsforscher richtet, ist im Vergleich zur gesprächstypologischen Variation die systematische Erfassung der sozialen und arealen Variation (vgl. Abschn. 3.2 und 3.3) dagegen zunächst nachrangig.

8. Ausblick

Der Aufbau eines nationalen Gesprächskorpus ist ein Projekt, das große Ressourcen und einen langen Atem erfordert. Während in der ersten Zeit die quantitative Erweiterung des Korpus und die Zusammenstellung einer Auswahl von Kommunikationsereignissen, die grundlegende

Merkmalskonstellationen von Gesprächssituationen prototypisch abdecken, im Vordergrund stehen, wird es erst nach vielen Jahren möglich und sinnvoll sein, so systematisch wie empirisch möglich die Lücken in der theoretisch und aufgrund vorliegender Forschungen anzunehmenden Gattungstaxonomie im Bestand mit Gattungsbeispielen zu füllen.

FOLK ist daher eine Langzeitaufgabe, die sich über mehrere Dekaden erstrecken wird. Der Aufbau eines solchen, für die linguistische Forschergemeinschaft und auch darüber hinaus unverzichtbaren Korpus ist nur an einer Institution wie dem Institut für Deutsche Sprache möglich. Der kontinuierliche Korpusausbau, die Weiterentwicklung der nötigen korpustechnologischen Instrumente und die Nachhaltigkeit der Archivierung kann nicht projektabhängig erfolgen, sondern aufgrund einer gesicherten grundständigen Finanzierung im Rahmen der institutionellen Aufgabenbestimmung, wissenschaftliche Infrastrukturen zu schaffen und dauerhaft zu gewährleisten. Zu einem solchen Korpusaufbaukonzept gehört für uns auch die Entwicklung von Instrumenten zur Korpuserstellung, die auch Nutzern außerhalb des IDS zur Verfügung gestellt werden. Dazu gehören insbesondere der Transkripteditor FOLKER (http://agd.ids-mannheim.de/html/folker.shtml) und ein Metadatendokumentationsschema sowie das Angebot, bei der Erstellung von Korpora nach den datentechnischen und den Transkriptions-Standards von FOLK beratend zur Seite zu stehen. Auf diese Weise kann die wissenschaftliche Gemeinschaft von unserem beim Aufbau von FOLK gewonnenen Know How nicht nur durch die in FOLK enthaltenen Daten, sondern auch durch die korpus-technologischen Instrumente, Verfahren und Kenntnisse profitieren. Außerdem entsteht so die Möglichkeit, dass auch außerhalb des IDS Daten gesammelt und aufbereitet werden können, die nach Abschluss ihrer Auswertung im Forschungsprojekt (zumindest teilweise) in FOLK integriert und damit der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden können. Wir sehen FOLK somit nicht nur als ein Projekt des IDS, sondern als ein kollaboratives Vorhaben, das durch die Mitarbeit von vielen Mitgliedern der wissenschaftlichen Gemeinschaft dieser so sehr nützt, wie es von ihr mitgetragen wird.

9. Literaturverzeichnis

ALLWOOD, JENS et al. (2000): The Spoken Language Corpus at the Linguistics Department, Göteborg University. In: Forum: Qualitative Research 1.3 (2000). URL: http://www.qualitative-research.net/index.php/fqs/article/view/1026

AUER, PETER (1992): Introduction: John Gumperz' Approach to Contextualization. In: AUER, PETER / DI LUZI, ALDO (Hgg.): The contextualization of language. Cambridge. 1–37.

BAUDE, OLIVIER et al. (2006): Corpus oraux. Paris.

BEREND, NINA (2005): Regionale Gebrauchsstandards – Gibt es sie und wie kann man sie beschreiben? In: EICHINGER, LUDWIG M. / KALLMEYER, WERNER (Hgg.): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? Berlin. 143–170.

BURNARD (2001): Where did we go wrong? A retrospective look at the design of the BNC. URL: http://users.ox.ac.uk/~lou/wip/silfitalk.html.

DEPPERMANN, ARNULF (2008): Verstehen im Gespräch. In: KÄMPER, HEIDRUN / EICHINGER, LUDWIG M. (Hgg.): Sprache – Kognition – Kultur. Sprache zwischen mentaler Struktur und kultureller Prägung. Berlin, New York. 225–261. (Jahrbuch des Instituts für deutsche Sprache).

DEPPERMANN, ARNULF / SPRANZ-FOGASY, THOMAS (2001): Aspekte und Merkmale der Gesprächssituation. In: BRINKER, KLAUS / ANTOS, GERD / HEINEMANN, WOLFGANG / SAGER, SVEN FREDERIK (Hgg.) (2001): Text- und Gesprächslinguistik. 2. Halbband. Berlin. 1148–1161.

ECKERT, PENELOPE (2000): Linguistic variation as social practice. Oxford.

FIEHLER, REINHARD / WAGENER, PETER (2005): Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. In: Gesprächsforschung 6 (2005). 136–147.
Online: http:// www.gespraechsforschung-ozs.de/heft2005/ heft2005.htm.

GLAS, REINHOLD / EHLICH, KONRAD (2000): Deutsche Transkripte 1950 bis 1995. Ein Repertorium. Hamburg.

GOFFMAN, ERVING (1963): Behavior in public places. Notes on the social organization of gatherings. New York.

GÜNTHNER, SUSANNE / KNOBLAUCH, HUBERT (1994): 'Forms are the food of faith'. Gattungen als Muster kommunikativen Handelns. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 4 (1994). 693–723.

HEATH, CHRISTIAN / VAN LEHN, DIRK (2008) : Construing interactivity: Enhancing engagement with new technologies in science centres and museums. In: Social Studies of Science 38 (2008). 63–96.

HEINEMANN, WOLFGANG / VIEHWEGER, DIETER (1991): Textlinguistik. Tübingen.

HENNE, HELMUT / REHBOCK, HELMUT (1982): Einführung in die Gesprächsanalyse. Berlin.

HITZLER, RONALD / BUCHER, THOMAS / NIEDERBACHER, ARNE (2005): Leben in Szenen. Formen jugendlicher Vergemeinschaftung heute. Opladen.

ISENBERG, HORST (1983): Grundfragen der Texttypologie. In: DANEŠ, FRANTIŠEK / VIEHWEGER, DIETER (Hgg.): Ebenen der Textstruktur. Berlin: Akademie der Wissenschaften der DDR. 303–342.

KALLMEYER, WERNER / KEIM, INKEN (2003): Eigenschaften von sozialen Stilen der Kommunikation: Am Beispiel einer türkischen Migrantinnengruppe. In: Osnabrücker Beiträge zur Sprachtheorie (OBST) 65 (2003). 35–56.

KOCH, PETER / OESTERREICHER, WULF (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36 (1985). 15–43.

KOTTHOFF, HELGA (1998): Spaß verstehen. Tübingen.

LEMNITZER, LOTHAR / ZINSMEISTER, HEIKE (2006): Korpuslinguistik. Eine Einführung. Tübingen.

LUCKMANN, THOMAS (1986): Grundformen der gesellschaftlichen Vermittlung des Wissens: Kommunikative Gattungen. In: Zeitschrift für Soziologie 27 (1986). 191–211.

MCCARTHY, MICHAEL / O'KEEFE, ANNE (2008) Corpora and spoken language. In: LÜDELING, ANKE / KYTÖ, MERJA (Hgg.): Korpuslinguistik/Corpus Linguistics. Berlin. 1008–1024.

MERKEL, SILKE / SCHMIDT, THOMAS (2009): Korpora gesprochener Sprache im Netz – eine Umschau. In: Gesprächsforschung 10 (2009). 70–93. http://www. gespraechsforschung-ozs.de/heft2009/px-merkel.pdf.

MONDADA, LORENZA (2007a): Imbrications de la technologie et de l'ordre interactionnel. L'organisation de vérifications et d'identifications de problèmes pendant la visioconférence. In: Réseaux 144. 141–182.

MONDADA, LORENZA (2007b): Turn Taking in multimodalen und multiaktionalen Kontexten. In: HAUSENDORF, HEIKO (Hg).: Gespräch als Prozess. Linguistische Aspekte der Zeitlichkeit verbaler Interaktion. Tübingen. 237–276.

MONDADA, LORENZA / SCHMITT, REINHOLD (2010) (Hgg.): Situationseröffnungen. Zur multimodalen Herstellung fokussierter Interaktion. Tübingen.

OOSTDIJK, NELLEKE (2002): The Design of the Spoken Dutch Corpus. In: PETERS, PAM / COLLINS, PETER / SMITH, ADAM (Hgg.): New Frontiers of Corpus Research. Amsterdam. 105–112.

SACKS, HARVEY (1978): Some technical considerations of a dirty joke. In: SCHENKEIN, JIM (Hg.): Studies in the organization of conversational interaction. New York. 249–269.

SACKS, HARVEY (1984): Notes on methodology. In: ATKINSON, JOHN MAXWELL / HERITAGE, JOHN (Hgg.): Structures of Social Action. Studies in Conversation Analysis. Cambridge, Mass. 21–27.

SCHEGLOFF, EMANUEL A. (1991): Reflections on talk and social structure. In: BODEN, DEIRDRE / ZIMMERMAN, DON H. (Hgg.) (1991): Talk and social structure. Cambridge MA. 44–71.

SCHMIDT, AXEL / NEUMANN-BRAUN, KLAUS (2004): Die Welt der Gothics – Spielraum düster konnotierter Transzendenz. Wiesbaden.

SCHMIDT, THOMAS (2005): Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze. In: Gesprächsforschung 6 (2005). 103–126. Online: http://www.gespraechsforschung-ozs.de/heft2005/heft2005.htm.

SCHMITT, REINHOLD (2007): Das Filmset als Arbeitsplatz. Multimodale Grundlagen einer komplexen Kooperationsform. In: TIITTULA, LIISA / PIITULAINEN, MARJA-LEENA / REUTER, EWALD (Hgg.): Die gemeinsame Herstellung professioneller Interaktion. Tübingen. 25–66.

SCHULZE, GERHARD (1992): Die Erlebnisgesellschaft: Kultursoziologie der Gegenwart. Frankfurt am Main.

STEGER, HUGO / DEUTRICH, HELGE / SCHANK, GERD / SCHÜTZ, EVA (1974): Redekonstellation, Redekonstellationstyp, Textexemplar, Textsorte im Rahmen eines Sprachverhaltensmodells. In: MOSER, HUGO (Hg.): Gesprochene Sprache. Jahrbuch 1972 des Instituts für Deutsche Sprache. Düsseldorf. 39–97.

THALER, VERENA (2007): Mündlichkeit, Schriftlichkeit, Synchronizität. Eine Analyse alter und neuer Konzepte zur Klassifizierung neuer Kommunikationsformen. In: Zeitschrift für germanistische Linguistik 35 (2007). 146–181.

Tikvah Kissmann, Ulrike (Hg.) (2009): Video Interaction Analysis: Methods and Methodology. Frankfurt am Main.

Wagener, Peter / Bausch, Karl-Heinz (Hgg.) (1997): Tonaufnahmen des gesprochenen Deutsch. Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven. Tübingen.

Wichman, Anne (2008): Speech corpora and spoken corpora. In: Lüdeling, Anke / Kytö, Merja (Hgg.): Korpuslinguistik/Corpus Linguistics. Berlin. 187–207.

10. Internetquellen der im Text aufgeführten Korpora

British National Korpus (BNC): http://www.natcorp.ox.ac.uk /corpus/index.xml

Corpus Gesproken Nederlands (CGN): http://lands.let.kun.nl/cgn/ehome.htm

Datenbank Gesprochenes Deutsch (DGD): http://agd.ids-mannheim.de/html/dgd.shtml;
Hamburger SFB 538 Mehrsprachigkeit: http://www.exmaralda.org/corpora/sfbkorpora.html

Deutsches Referenzkorpus am Institut für Deutsche Sprache Mannheim: http://www.ids-mannheim.de/kl/projekte/korpora

Deutsches Wörterbuch der Gegenwartssprache (DWDS) der Berlin Brandenburgischen Akademie der Wissenschaften: http:// www.dwds.de/textbasis

Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK): http://agd.ids-mannheim.de/html/folk.shtml

Göteborg Spoken Language Corpus (GSLC): http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

Korpussammlung der Universität Leipzig: http://corpora.informatik.uni-leipzig.de/download.html